BLIND SOURCE SEPARATION OF FIR CONVOLUTIVE MIXTURES: APPLICATION TO SPEECH SIGNALS

Cédric Févotte, Alexandra Debiolles, Christian Doncarli

Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN)
UMR CNRS 6597, 1 rue de la Noë, BP 92101, 44321 Nantes Cedex 03, France
fevotte@irccyn.ec-nantes.fr

ABSTRACT

In this paper we present a simple method to deal with Blind Source Separation (BSS) of Finite Impulse Response (FIR) convolutive mixtures. The global method proceeds in two steps. The first step consists in *separating each source contribution* in the mixture. This step provides *several filtered version of each source*. The second step consists in retrieving the original sources from the set of filtered versions of each source using a *blind system identification* method. We present some results on a mixture of speech and music.

1. INTRODUCTION

Blind Source Separation has many applications in Audio Signal Processing (see [1] for an overview of Audio Source Separation applications). Usually we have to deal with convolutive mixtures, for example to take into account the reverberation in a room.

In this paper we aim at presenting how the separation of a Finite Impulse Response (FIR) convolutive mixture can be tackled with the use of a (joint) block-diagonalization procedure. The method proceeds in two steps.

First step consists in separating each source contribution to the mixture. The convolutive mixing is rearranged into a multiplicative mixing of new sources introducing proper variables. But some of these new sources are dependent. However, the contributions of the different sources can be separated by using the standard BSS algorithm SOBI extended for the particular case when some of the sources are dependent. However, the algorithm do not provide the original sources but *several filtered version of each source*.

Then, the second step consists in recovering the original sources from the set a of filtered versions of each source obtained from the first step. In the literature, this problem is usually named *blind system identification* or *blind deconvolution*.

Each step of the global BSS method was published separately by several authors. The first step (extension of SOBI to convolutive mixtures) is presented in [2]. The algorithm is based on the joint block-diagonalization of several whitened covariance matrices. The method assumes that the sources are stationary but still performs well on audio signals as long as they are uncorrelated. Some time-frequency approaches were proposed by several authors to deal with non-stationary signals but we will not address such a level of generality here (see [3] for an overview of blind separation methods for convolutive mixtures using block-diagonalization).

The second step has been widely investigated in the literature. An overview of *blind system identification* (BSI) methods is available in [4]. We will briefly describe the method based on subspace decomposition presented in [5].

The only novelty in this paper is the combination of the first step (SOBI convolutive) and the second step (BSI) to perform complete separation instead of partial separation as in [2].

In Section 2 we introduce assumptions and notations, and we show how the convolutive mixing can be turned into an instantaneous mixing. In Sections 3 and 4 we present briefly the source contributions separation step and the blind identification step. An illustration of the performance of the method over a mixture of speech and guitar is presented in Section 5.

2. BACKGROUND

2.1. Aim and assumptions

We consider the following discrete-time noiseless FIR MIMO model:

$$\mathbf{x}[t] = \mathbf{H}[0]\,\mathbf{s}[t] + \mathbf{H}[1]\,\mathbf{s}[t-1] + \ldots + \mathbf{H}[L]\,\mathbf{s}[t-L] \quad (1)$$

where $\mathbf{x}[t] = [x_1[t], \ldots, x_m[t]]^T$ is the vector of size m containing the observations, $\mathbf{s}[t] = [s_1[t], \ldots, s_n[t]]^T$ is the vector of size n containing the sources (assumed zero-mean and mutually uncorrelated at every time instant), $\mathbf{H}[k] = \{h_{ij}[k]\}$, $k = 0 \ldots L$, are $m \times n$ matrices with $m > n$.

The overall objective of BSS is to obtain estimates of the mixing filters and/or estimates of the sources up to standard BSS indeterminacies on ordering, scale and phase.

2.2. Back to instantaneous mixing

We recall from [2] how the convolutive mixing (1) can be rearranged into an instantaneous mixing.

2.3. Notations

Let L' be an integer such that $mL' \geq n(L+L')$ (L' exists when $m > n$). We note, for $i = 1, \ldots, n$:

$$\mathbf{S}_i[t] = [s_i[t], \ldots, s_i[t-(L+L')+1]]^T$$

and for $j = 1, \ldots, m$

$$\mathbf{X}_j[t] = [x_j[t], \ldots, x_j[t-L'+1]]^T$$

where $\cdot^T$ denotes "transpose". Then we introduce:

$$\begin{aligned} \mathbf{S}[t] &= \left[\mathbf{S}_1[t]^T, \ldots, \mathbf{S}_n[t]^T\right]^T \\ \mathbf{X}[t] &= \left[\mathbf{X}_1[t]^T, \ldots, \mathbf{X}_m[t]^T\right]^T \end{aligned}$$

$\forall t$, $\mathbf{S}[t]$ is a column vector of size $n(L+L')$ and $\mathbf{X}[t]$ is a column vector of size mL'. For simplicity we note $N = n(L+L')$ and $M = mL'$.

For $i = 1, \ldots, n$ and $j = 1, \ldots, m$ we note $\mathbf{A}_{ij}$ the following $L' \times (L+L')$ Sylvester matrix:

$$\mathbf{A}_{ij} = \begin{bmatrix} h_{ij}[0] & \ldots & h_{ij}[L] & 0 & \ldots & 0 \\ & \ddots & & \ddots & & \\ & & \ddots & & \ddots & \\ 0 & \ldots & 0 & h_{ij}[0] & \ldots & h_{ij}[L] \end{bmatrix}$$

Finally, we note:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \ldots & \mathbf{A}_{1n} \\ \vdots & & \vdots \\ \mathbf{A}_{1m} & \ldots & \mathbf{A}_{mn} \end{bmatrix}$$

$\mathbf{A}$ is a $M \times N$ matrix which satisfies:

$$\mathbf{X}[t] = \mathbf{A}\,\mathbf{S}[t] \tag{2}$$

In the following we assume that $\mathbf{A}$ is full rank.

Eq. (2) shows that the convolutive mixing (1) can be written as an instantaneous mixture. Such mixtures (2) have been widely studied in BSS/ICA literature. However the big difference here is that the components of $\mathbf{S}[t]$ are not *all* mutually independent: when the sources are not white, for $i = 1, \ldots, n$, the components of $\mathbf{S}_i[t]$ are dependent.

3. STEP 1: SEPARATION OF THE CONTRIBUTIONS OF EACH SOURCE

In this section we briefly describe the source contributions separation step. We will assume that the sources are stationary. For a more general study, in particular in a noisy environment, the reader should refer to [3] and references therein concerning A. Belouchrani and K. Abed-Meraim work on the topic. The key of the method is to formulate the overall problem described by Eq. (2) in the time-lag plane.

For $(t, \tau) \in \mathbb{Z}^2$ we note $\mathcal{R}_{\mathbf{SS}}[t, \tau]$ the covariance matrix of $\mathbf{S}[t]$ defined by:

$$\mathcal{R}_{\mathbf{SS}}[t, \tau] \overset{\text{def}}{=} \mathrm{E}\{\mathbf{S}[t]\,\mathbf{S}[t+\tau]^H\}$$

Since the sources are assumed stationary we have:

$$\mathcal{R}_{\mathbf{SS}}[t, \tau] = \mathcal{R}_{\mathbf{SS}}[\tau] \tag{3}$$

The vector signals $\mathbf{S}_1[t], \ldots, \mathbf{S}_n[t]$ being mutually uncorrelated, the $N \times N$ covariance matrix $\mathcal{R}_{\mathbf{SS}}[\tau]$ is block-diagonal with n blocks of dimensions $(L+L')$, such that:

$$\mathcal{R}_{\mathbf{SS}}[\tau] = \begin{bmatrix} \mathcal{R}_{\mathbf{s}_1\mathbf{s}_1}[\tau] & & \\ & \ddots & \\ & & \mathcal{R}_{\mathbf{s}_n\mathbf{s}_n}[\tau] \end{bmatrix}$$

With Eq. (2) we have:

$$\mathcal{R}_{\mathbf{XX}}[\tau] = \mathbf{A}\,\mathcal{R}_{\mathbf{SS}}[\tau]\,\mathbf{A}^H \tag{4}$$

3.1. Generalization of SOBI

A two-steps separation method (whitening and rotation) can be devised from (4) [2, 3].

3.1.1. Whitening

In the FIR convolutive case, whitening consists in finding a matrix $\mathbf{W}$ of dimensions $N \times M$ such that:

$$\mathbf{W}\,\mathbf{A}\,\mathbf{B}\,\mathbf{A}^H\,\mathbf{W}^H = \mathbf{I}_M \tag{5}$$

where $\mathbf{B}$ is a $N \times N$ bloc-diagonal matrix positive definite with n blocks of dimension $(L+L')$. In practice, using $\mathbf{B} = \mathcal{R}_{\mathbf{SS}}[0]$, $\mathbf{W}$ can be computed from the eigenelements of the following estimation of $\mathcal{R}_{\mathbf{XX}}[0]$ [3]:

$$\begin{aligned} \hat{\mathcal{R}}_{\mathbf{XX}}[0] &\overset{\text{def}}{=} \frac{1}{T}\sum \mathbf{X}[t]\,\mathbf{X}[t]^H && (6) \\ &\approx \mathbf{A}\,\hat{\mathcal{R}}_{\mathbf{SS}}[0]\,\mathbf{A}^H && (7) \end{aligned}$$

3.1.2. Rotation

The second step of the method is the estimation of $\mathbf{U} = \mathbf{W}\,\mathbf{A}\,\mathbf{B}^{\frac{1}{2}}$. It is shown in Section 3.1.3 that some estimates of the sources can be retrieved from $\mathbf{W}$ and $\mathbf{U}$.

Let us define the following "whitened" covariance matrices:

$$\underline{\mathcal{R}}_{\mathbf{XX}}[\tau] = \mathbf{W}\,\mathcal{R}_{\mathbf{XX}}[\tau]\,\mathbf{W}^H \tag{8}$$

With Eq. (4) we have:

$$\begin{aligned}\underline{\mathcal{R}}_{\mathbf{XX}}[\tau] &= \mathbf{W}\,\mathbf{A}\,\mathcal{R}_{\mathbf{SS}}[\tau]\,\mathbf{A}^H\,\mathbf{W}^H \\ &= \mathbf{U}\,(\mathbf{B}^{-\frac{1}{2}}\,\mathcal{R}_{\mathbf{SS}}[\tau]\,\mathbf{B}^{-\frac{H}{2}})\,\mathbf{U}^H \end{aligned} \quad (9)$$

Since $\mathbf{B}^{-\frac{1}{2}}$, $\mathbf{B}^{-\frac{H}{2}}$ and $\mathcal{R}_{\mathbf{SS}}[\tau]$ are block-diagonal matrices we see that $\mathbf{U}$ block-diagonalizes $\underline{\mathcal{R}}_{\mathbf{XX}}[\tau]$ for all τ.

Thus, $\mathbf{U}$ can be retrieved in theory from the block-diagonalization of any matrix $\underline{\mathcal{R}}_{\mathbf{XX}}[\tau]$. In practice an estimate of $\mathbf{U}$ should rather be computed from the joint block-diagonalization (JBD) of a set of K matrices $\{\underline{\mathcal{R}}_{\mathbf{XX}}[\tau_i], i = 1...K\}$. JBD provides a more robust estimate of $\mathbf{U}$ with respect to estimation errors on $\underline{\mathcal{R}}_{\mathbf{XX}}[\tau]$ and reduces indeterminacies in the same way joint-diagonalization does [6]. JBD provides a matrix $\mathbf{U}_{JBD}$ such that:

$$\mathbf{U}_{JBD} = \mathbf{U}\,\mathbf{P} \quad (10)$$

where $\mathbf{P}$ is a $N \times N$ unitary matrix that models JBD indeterminacies. $\mathbf{P}$ is the product of a block-diagonal unitary matrix with n blocks of dimension $(L + L') \times (L + L')$ with a permutation matrix of these blocks. A Jacobi-like JBD algorithm is presented in [7].

3.1.3. Retrieving the sources

In this section we compute estimates of the sources (up to unknown filters) from $\mathbf{U}_{JBD}$ and $\mathbf{W}$. We define the following column vector $\hat{\mathbf{S}}[t]$ of dimension N:

$$\hat{\mathbf{S}}[t] = \mathbf{U}_{JBD}^H\,\mathbf{W}\,\mathbf{X}[t] \quad (11)$$

Eq.'s (10) and (2) yield:

$$\hat{\mathbf{S}}[t] = \mathbf{C}\,\mathbf{S}[t] \quad (12)$$

with:

$$\mathbf{C} \overset{\text{def}}{=} \mathbf{P}^H\,\mathbf{B}^{-\frac{1}{2}} \quad (13)$$

$\mathbf{C}$ is a $N \times N$ block-diagonal matrix with n blocks $\mathbf{C}_1, \ldots, \mathbf{C}_n$ of dimensions $(L + L') \times (L + L')$. We decompose $\hat{\mathbf{S}}[t]$ into n sub-vectors of dimension $(L + L')$ such that

$$\hat{\mathbf{S}}[t] = \left[\hat{\mathbf{S}}_1[t]^T, \ldots, \hat{\mathbf{S}}_n[t]^T\right]^T$$

Then, for $i = 1, \ldots, n$, we have:

$$\hat{\mathbf{S}}_i[t] = \mathbf{C}_i\,\mathbf{S}_i[t] \quad (14)$$

We recall that $\mathbf{S}_i[t] = [s_i[t], \ldots, s_i[t - (L + L') + 1]]^T$. Hence, Eq. (14) means that each component of $\hat{\mathbf{S}}_i[t]$ is a FIR filtered version of the i^{th} source $s_i[t]$. The coefficients of the filters are contained in corresponding rows of $\mathbf{C}_i$. Then, for each source $s_i[t]$, we retrieve $(L + L')$ filtered versions of $s_i[t]$. Thus, a further blind SIMO system identification step is required to estimate the original sources instead of filtered versions of them.

4. STEP 2: BLIND IDENTIFICATION

We now shortly describe the subspace method presented in [5] to handle the deconvolution of the several filtered versions of each source we obtained from the first step. Let us consider the deconvolution problem of a single source $d[t] = s_i[t]$. The deconvolution problem expressed by Eq. (14) matches the following structure:

$$\begin{aligned} y_1[t] &= f_1[0]\,d[t] + \ldots + f_1[Q]\,d[t - Q] \\ &\vdots \\ y_P[t] &= f_P[0]\,d[t] + \ldots + f_P[Q]\,d[t - Q] \end{aligned}$$

where $P = L + L'$, $Q = L + L' - 1$, $y_k[t]$ is the k^{th} entry of $\hat{\mathbf{S}}_i[t]$ and $[f_k[0], \ldots, f_k[Q]]$ is the k^{th} row of $\mathbf{F}_i$. Our goal is to estimate the filters parameters, that we stack in:

$$\mathbf{f} = [f_1[0], \ldots, f_1[Q], \ldots, f_P[0], \ldots, f_P[Q]]^T \quad (15)$$

Provided the filters parameters we will be able to recover the sources with inverse filtering.

Let $\mathbf{W}$ be an integer "window parameter". We define, $\forall i = 1, \ldots, P$:

$$\mathbf{y}_i[t] = [y_i[t], \ldots, y_i[t - W + 1]]^T$$

and:

$$\begin{aligned} \mathbf{y}[t] &= \left[\mathbf{y}_1[t]^T, \ldots, \mathbf{y}_P[t]^T\right]^T \\ \mathbf{d}[t] &= [d[t], \ldots, d[t - W - Q + 1]]^T \end{aligned}$$

We define $\forall i = 1, \ldots, P$:

$$\mathbf{F}_W^{(i)} = \begin{bmatrix} f_i[0] & \ldots & f_i[Q] & 0 & \ldots & 0 \\ & \ddots & & \ddots & & \\ & & \ddots & & \ddots & \\ 0 & \ldots & 0 & f_i[0] & \ldots & f_i[Q] \end{bmatrix}$$

(size $W \times (Q + W)$) and:

$$\mathbf{F}_W = \begin{bmatrix} \mathbf{F}_W^{(1)} \\ \vdots \\ \mathbf{F}_W^{(P)} \end{bmatrix} \quad (16)$$

(size $PW \times (Q + W)$). With these notations we have:

$$\mathbf{y}[t] = \mathbf{F}_W\,\mathbf{d}[t] \quad (17)$$

The parameter W must chosen such as $PW \geq (Q + W)$, which means that the system in Eq. (17) becomes overdetermined. The key theorem of [5] is that, if $W \geq Q$ and if $\mathbf{F}_{W-1}$ is full column rank, then the range of the columns of $\mathbf{F}_W$ uniquely determines $\mathbf{f}$. This means that if we can determine the range of $\mathbf{F}_W$, we only have to compute a basis of it in the Sylvester matrix form of $\mathbf{F}_W$ to recover $\mathbf{f}$ (up to scalar factor).

From Eq. (17), we have:

$$\mathcal{R}_{\mathbf{yy}}[0] = \mathbf{F}_W \, \mathcal{R}_{\mathbf{dd}}[0] \, \mathbf{F}_W^H \qquad (18)$$

If $\mathbf{F}_W$ and $\mathcal{R}_{\mathbf{dd}}[0]$ are full column rank matrices, $\mathcal{R}_{\mathbf{yy}}[0]$ is a matrix of rank $W+Q$ and the range of $\mathbf{F}_W$ is simply the space orthogonal to the *null subspace* (or *noise subspace* in the presence of noise) of $\mathcal{R}_{\mathbf{yy}}[0]$. The *null subspace* is the range of the $PW-Q-W$ eigenvectors of $\mathcal{R}_{\mathbf{yy}}[0]$ associated to the eigenvalue 0. If $\mathbf{E}_W$ denotes the $PW \times PW - Q - W$ matrix containing these eigenvectors, $\mathbf{f}$ can be simply estimated as the minimizer of:

$$q(\mathbf{f}) = \|\mathbf{E}_W^H \, \mathbf{F}_W\|_F \qquad (19)$$

With proper variables, $q(\mathbf{f})$ can expressed as a quadratic form in $\mathbf{f}$ and its minimization under the constraint $\|\mathbf{f}\|_F = 1$ thus amounts to the computation of an eigenvector. See [5] for full details.

5. SIMULATIONS RESULTS

We present some results on a noiseless mixture of two sources (one is speech, the other is electric guitar). The matrix of mixing filters is arbitrarily chosen as:

$$\mathbf{H}[z] = \begin{bmatrix} 1+0.8z^{-1}+0.5z^{-2} & 0.8+0.7z^{-1}+0.4z^{-2} \\ 0.9+0.4z^{-1}+0.6z^{-2} & 1+0.9z^{-1}+0.3z^{-2} \\ 0.7+0.6z^{-1}+0.5z^{-2} & 0.8+0.3z^{-1}+0.6z^{-2} \end{bmatrix}$$

The sources, mixtures and estimated sources are presented on Fig. 1. We computed the source separation criteria described in [8]. The Source to Distortion Ratio (SDR) measures the global error made on the estimates of the sources, whereas the Source to Interference Ratio (SIR) only measures the contribution of other sources in the estimation of one source and the Source to Artifacts Ratio (SAR) only measures the proportion of Artifacts due to the algorithm in the estimates. The figures show high quality source separation:

	Speech	Guitar
SDR (dB)	56.6	41.12
SIR (dB)	83.5	41.9
SAR (dB)	56.07	49.0

6. CONCLUSION

The advantage of the global method we described is that it allows *complete separation* whereas many BSS methods dealing with convolutive mixtures only provide *partial separation*, that is, sources estimated only up to a filter.

But this is at the price of heavy computation loads, which is the main disadvantage of the method. The results

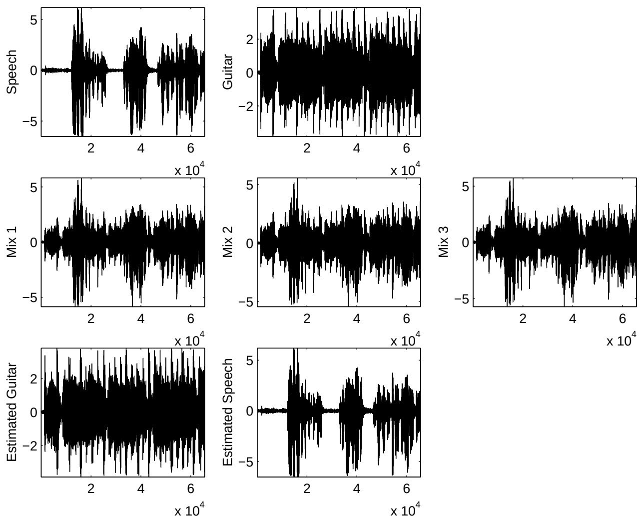

Fig. 1. Evaluation of extended SOBI + BSI on a mixture of speech and guitar

we obtained on short length filters happened to be very good up to $L = 6$. The whole method strongly relies on the joint block-diagonalization procedure which happened to fail with longer filters.

7. REFERENCES

[1] E. Vincent, C. Févotte, R. Gribonval, and al, "A tentative typology of audio source separation tasks," in 4th *Symposium on Independent Component Analysis and Blind Source Separation (ICA'03)*, Nara, Japan, 2003.

[2] H. Bousbiah-Salah, A. Belouchrani, and K. Abed-Meraim, "Jacobi-like algorithm for blind signal separation of convolutive mixtures," *Electronics Letters*, vol. 37, no. 16, pp. 1049–1050, Aug 2001.

[3] C. Févotte and C. Doncarli, "A unified presentation of blind source separation methods for convolutive mixtures using block-diagonalization," in *Proc. 4th Symp. on Independent Component Analysis and Blind Source Separation*, 2003.

[4] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1310–1322, Aug. 1997.

[5] E. Moulines, P. Duhamel, J-F. Cardoso, and S. Mayrargue, "Susbspace methods for the blind identification of multichannel fir filters," *IEEE Trans. Signal Processing*, vol. 43, no. 2, Feb 1995.

[6] A. Belouchrani, K. Abed-Meraim, J. F. Cardoso, and É. Moulines, "A blind source separation technique based on second order statistics," *IEEE Trans. Signal Processing*, vol. 45, no. 2, pp. 434–444, Feb 1997.

[7] A. Belouchrani, K. Abed-Meraim, and Y. Hua, "Jacobi-like algorithms for joint block diagonalization: Application to source localization," in *Proc. ISPACS*, Nov. 1998.

[8] E. Vincent R. Gribonval, L. Benaroya and C. Févotte, "Proposals for performance measurement in source separation," in 4th *Symposium on Independent Component Analysis and Blind Source Separation (ICA'03)*, Nara, Japan, 2003.